# Data Mining & Statistical Learning

Statistics 640/444 • Fall 2014 • Tuesdays & Thursdays 4:00 - 5:15pm

**Instructor**:
Genevera Allen
Office: Duncan Hall 2098.
Office Hour: Thursday 5:15 - 6:15pm
Email: gallen@rice.edu

**Teaching Assistant**:
John Nagorski
Office: Duncan Hall 1038.
Office Hour: Friday 12:00 - 1:00pm
Email: john.nagorski@rice.edu

**Time & Location**:
Class: Tuesdays & Thursdays 4:00 - 5:15pm. Location: Herzstein Hall 210.
Help Session: Wednesdays 11:00 - 12:00pm. Location: Duncan Hall 3076.

**Course Description**: This course is a survey of statistical learning methods. The course will cover major statistical learning methods and concepts for both supervised and unsupervised learning. Topics covered include penalized regression and classification, support vector machines, kernel methods, model selection, matrix factorizations and completion, graphical models, clustering, boosting and ensemble learning.

**Course Objectives & Outcomes**: Students will learn how and when to apply statistical learning techniques, their comparative strengths and weaknesses, and how to critically evaluate the performance of learning algorithms. Students completing this course should be able to (i) apply basic statistical learning methods to build predictive models, (ii) properly tune and select statistical learning models, (iii) correctly assess model fit and error, and (iv) build an ensemble of learning algorithms.

**Stat 640 vs. 444**: Students registered in 444 and 640 will have different homework assignments, different midterm exams, and be graded by different criteria for the data mining competition. Stat 444 will focus on application and interpretation of statistical learning methods. Stat 640 will additionally focus on how these methods work, how they are implemented, and their mathematical properties. Students registered in Stat 444 are strongly encouraged to attend the extra help sessions.

**Prerequisites**: Linear Algebra & Scientific Programming in a language such as `R`, `Matlab`, or `Python`. Recommended: Linear Regression & Mathematical Probability and Statistics. Please speak to the instructor if you have any concerns regarding prerequisites.

**Recommended Textbooks**: Elements of Statistical Learning by Hastie, Tibshirani & Friedman; available online at `http://www-stat.stanford.edu/~tibs/ElemStatLearn`.
(640) Statistics for High-Dimensional Data by Buhlmann & van de Geer.
(444) Introduction to Statistical Learning by James, Witten, Tibshirani & Hastie; available online at `http://www-bcf.usc.edu/~gareth/ISL/`.

**Course Webpage**: `http://www.stat.rice.edu/~gallen/stat640.html`. Please check the course webpage frequently for announcements and homework assignments.

**Grading Policy**:

| | |
|---|---|
| Homeworks | 25% |
| Mid-Term Exam | 25% |
| Competition & Reports | 45% |
| Class Participation | 5% |

**Registrar Deadlines**:

| | |
|---|---|
| Friday, September 5 | Add Deadline |
| Friday, October 10 | Drop Deadline |

**Attendance Policy**: Student are encouraged, but not required to attend lectures.

**Homeworks**: There will be four homework assignments consisting of (640) mathematical proofs and derivations and (640 & 444) data analysis. Approximate dates:

| | Homework Assigned | Homework Due |
|---|---|---|
| Homework I | September 4 | September 18 |
| Homework II | September 23 | October 9 |
| Homework III | October 16 | October 30 |
| Homework IV | November 11 | November 25 |

A hard copy of homeworks are due at 5:15pm and can be turned in at class or to the TA in Duncan Hall on the due date. Late homeworks will NOT be accepted, NO exceptions. Homeworks may be discussed with classmates but must be written and submitted individually.

**Midterm Exam**: There will be a take-home midterm exam assigned on **November 6** and due on **November 13**. The exam will be three hours long and is open books and open notes. Internet use is prohibited and the exam must be completed individually according to the honor code.

**Data Mining Competition**: As part of the class, students will compete against each other in a data mining contest. The competition will begin on **Friday, September 5** and can be completed in teams of two people. Grades will be based upon two progress reports and a final report (one per team) as well as the contest results. The winning teams for 444 and 640 in terms of prediction accuracy, the team with the most innovative solution (640), and the team with the most interesting finding (444) will automatically receive the highest grades. Further details about the contest along with specific grading criteria will be given in a separate document and discussed in class.

| | |
|---|---|
| Contest Opens | 12:00am September 5 |
| Progress Report I | 5:15pm October 2 |
| Progress Report II | 5:15pm November 4 |
| Contest Closes | 11:59pm December 1 |
| In-Class Contest Presentations | December 2 & 4 |
| Final Report Due | 5:00pm December 5 |

**Tentative Schedule**: (Also see separate schedule document).

| | |
|---|---|
| August 26 | Introduction |
| August 28 | Review & KNN |
| September 2 | Least Squares & Ridge |
| September 4 | Ridge, PCA & PLS Regression |
| September 9 | Sparse Regression I |
| September 11 | Sparse Regression II |
| September 16 | Linear Discriminant Analysis |
| September 18 | LDA & Logistic Regression |
| September 23 | Sparse Classification |
| September 25 | Support Vector Machines I |
| September 30 | SVMs II |
| October 2 | Kernel Methods |
| October 7 | Model Selection I |
| October 9 | Model Selection II |
| October 16 | Recap: Supervised Learning |
| October 21 | Matrix Factorizations I |
| October 23 | Matrix Factorizations II |
| October 28 | Graphical Models I |
| October 30 | Graphical Models II |
| November 4 | Clustering I |
| November 6 | Clustering II |
| November 11 | CART |
| November 13 | Ensemble Learning |
| November 18 | Boosting I |
| November 20 | Boosting & Random Forests |
| November 25 | Random Forests |
| December 2 | Competition Presentations |
| December 4 | Competition Presentations |

Students with a documented disability requiring accommodations should speak with the instructor during the first two weeks of class.

This syllabus is subject to change with reasonable advance notice by the instructor.